Áp Dụng Mô Hình ANN Để :
PHÂN LOẠI UNG THƯ SỬ DỤNG DỮ LIỆU BIỂU HIỆN GEN

Nhóm Bảo Vệ:
Tạ Văn Nhân
Nguyễn Thành Trung

VNU - HUS | 2019

Nội Dung

1 Giới thiệu

- Bệnh bạch cầu là một loại ung thư máu gồm có hai loại: lymphoblastic cấp tính (ALL), bạch cầu tủy cấp tính (AML). Xác định và phân loại bệnh bạch cầu là rất cần thiết bởi việc điều trị thay đổi đáng kể theo tiểu loại của bệnh bạch cầu.
- Thông thường phương pháp phân loại ung thư dựa trên hình thái đặc điểm đã được xác định là không đầy đủ. Các tế bào có thể giống nhau về mặt hình thái nhưng phản ứng rất trái ngược với thuốc và phương pháp điều trị. Hồ sơ biểu hiện gen của các tế bào cung cấp thông tin hữu ích để phân loại bệnh.
- Mô hình ANN kết hợp với phương pháp giảm chiều dữ liệu PLS được áp dụng trong dữ liệu biểu hiện gen để phân loại ung thư bạch cầu một cách hiệu quả.

2 Dữ Liệu và Phương Pháp

2.1 Dữ Liệu

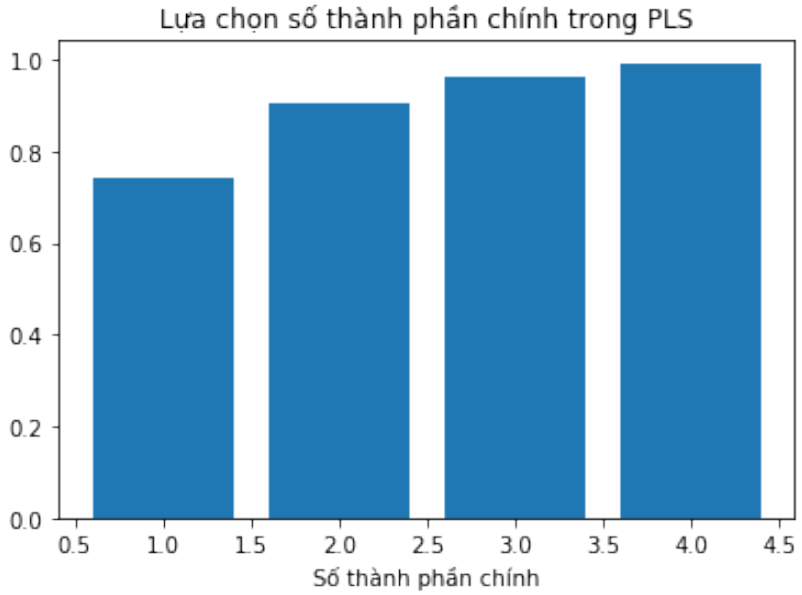
- Dữ liệu biểu hiện gen trong nghiên cứu của Golub và các cộng sự vào năm 1999, dữ liệu này bao gồm 72 mẫu bệnh bạch cầu trong trong đó có 47 bệnh bạch cầu lymphoblastic cấp tính (ALL) mẫu và 25 mẫu bệnh bạch cầu myleloid cấp tính (AML).
- Tập huấn luyện có 38 mẫu bệnh, tập kiểm thử có 34 mẫu bệnh. Mỗi mẫu được tạo thành từ hồ sơ biểu hiện gen của 7129 gen.

2 Dữ Liệu và Phương Pháp

2.2 Phương pháp

2.2.1 Partial Least Squares

Partial Least Squares (PLS) Regression là phương pháp giảm chiều dữ liệu hướng đến tìm một siêu phẳng mà hình chiếu các phần tử trong tích chéo của ma trận đầu vào X và ma trận đầu ra Y lên siêu phẳng đó có phương sai lớn nhất.



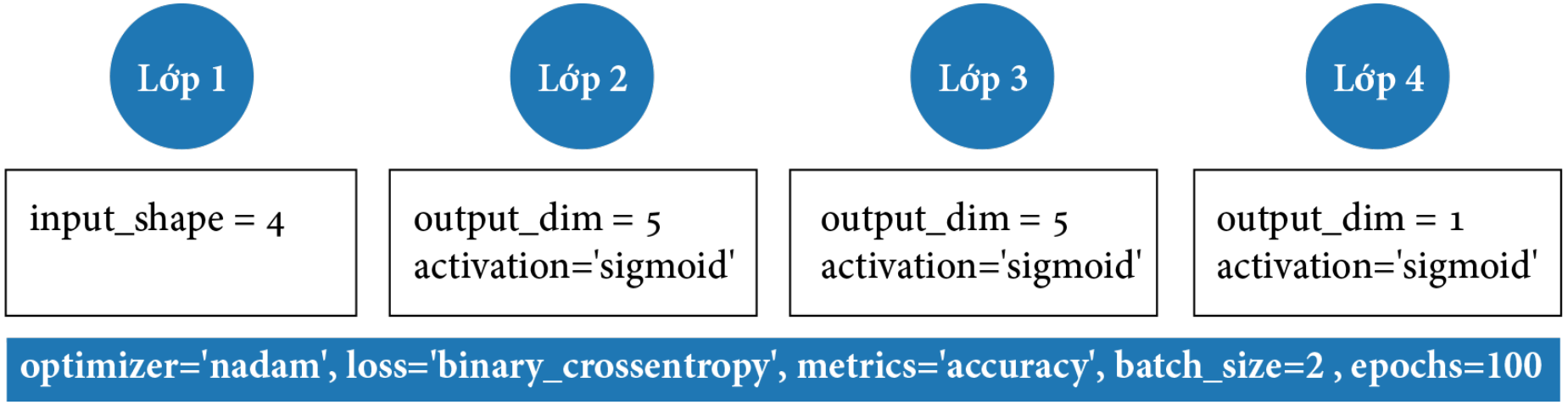
Comp 1	Comp 2	Comp 3	Comp 4
-10.621	-24.308	4.717	0.725
-6.113	0.702	-17.932	-28.236
-6.427	-43.662	21.911	2.57
-6.353	-23.285	0.963	4.896
-28.272	24.792	-0.206	-8.363
...	...	...	...

2 Dữ Liệu và Phương Pháp

2.2 Phương pháp

2.2.2 Artificial neural network

- Các hàm kích hoạt được sử dụng đều là hàm sigmoid.
- Phương pháp tối ưu hóa là NADAM, batch size bằng 2 và epochs bằng 100.
- Hàm tổn thất được sử dụng là hàm binary-crossentropy



2 Dữ Liệu và Phương Pháp

2.2 Phương pháp

2.2.3 Một số mô hình máy học

- SVM: tối ưu hóa lề giữa hai lớp dữ liệu, sử dụng các hàm kernel để tăng chiều dữ liệu khi chưa có sự phân tách tốt giữa hai lớp.
- K-means clustering: phân cụm dữ liệu dựa trên việc tối ưu hóa dần các cụm.
- Logistic regression: có thể coi như một mô hình ANN đơn giản trong đó chỉ có hai lớp, lớp đầu vào và lớp đầu ra, hàm kích hoạt được sử dụng là hàm sigmoid.
- Ngoài ra còn các mô hình máy học khác như Naive Bayes, Random forest, XG boots.

3 Kết Quả

3.1 So sánh giữa PCA và PLS

- Số thành phần chính của PCA và PLS đều được lựa chọn để giải thích trên 99% dữ liệu, PCA chọn được 36 thành phần chính, PLS chọn được 4 thành phần chính.
- Hầu hết các mô hình khi sử dụng PLS cho độ chính xác cao hơn khi sử dụng PCA, duy nhất mô hình SVM cho độ chính thì ngược lại.

PCA

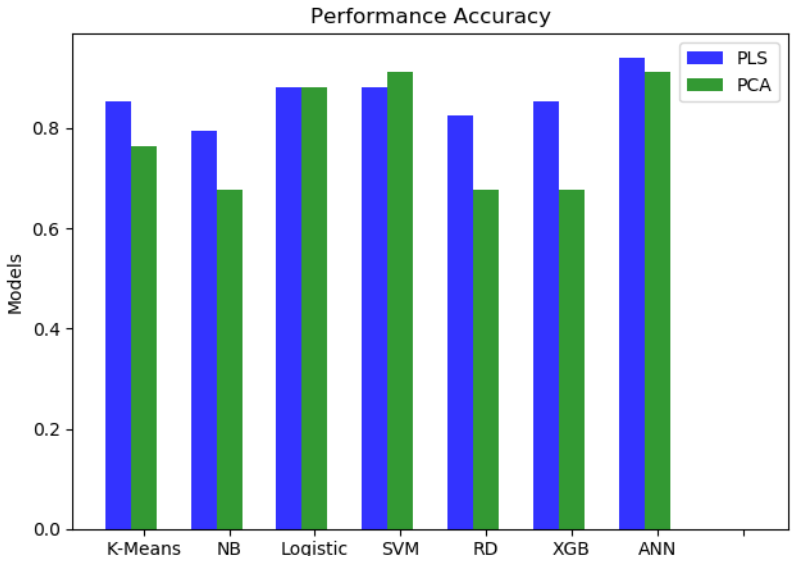
	Accuracy	Sensitivity	Precisions	F-measure
K-Means	0.765	0.76470588	0.7901584	0.7479656
Naive Bayes	0.676	0.67647059	0.7674959	0.6688548
Logistic	0.882	0.88235294	0.8823529	0.8823529
SVM	0.912	0.91176471	0.9232737	0.9095486
Random Forest	0.676	0.67647059	0.6804954	0.6779009
XG Boots	0.676	0.67647059	0.7912713	0.6066897
ANN	0.912	0.91176471	0.9141383	0.9121548

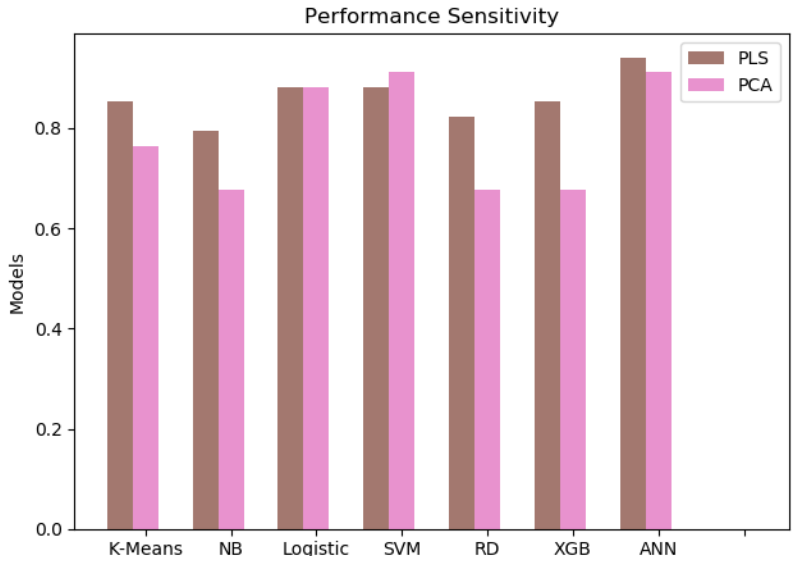
PLS

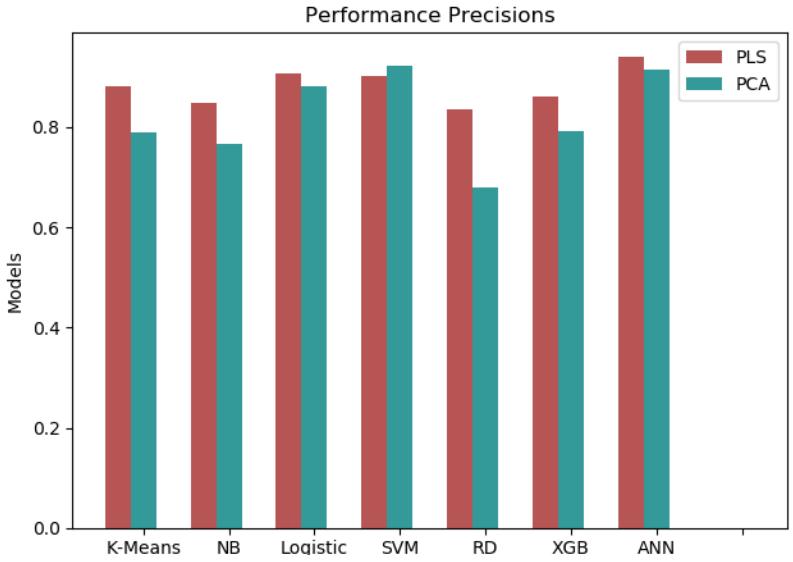
	Accuracy	Sensitivity	Precisions	F-measure
K-Means	0.853	0.85294118	0.88235294	0.8451265
Naive Bayes	0.794	0.79411765	0.84749455	0.7751356
Logistic	0.882	0.88235294	0.90849673	0.8831699
SVM	0.882	0.88235294	0.90196078	0.8778966
Random Forest	0.824	0.82352941	0.83627451	0.8168449
XG Boots	0.853	0.85294118	0.86026505	0.8492476
ANN	0.941	0.94117647	0.94117647	0.9411765

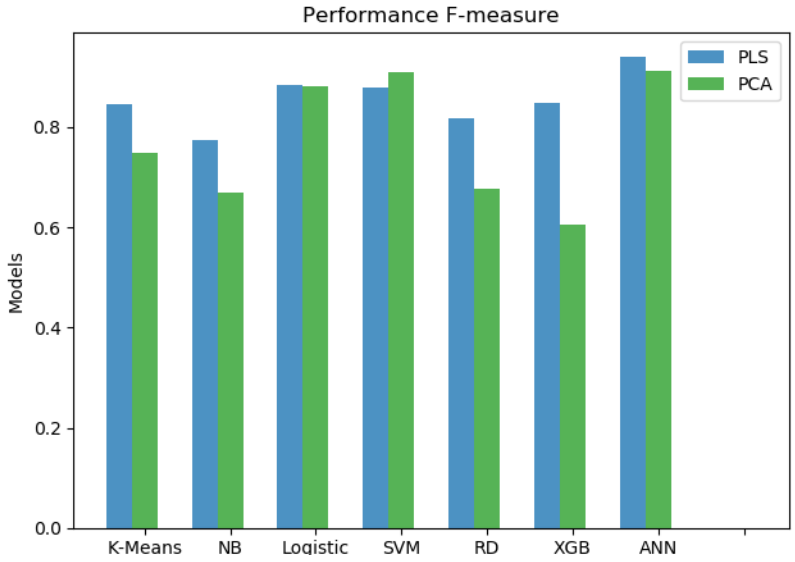
3 Kết Quả

3.1 So sánh giữa PCA và PLS





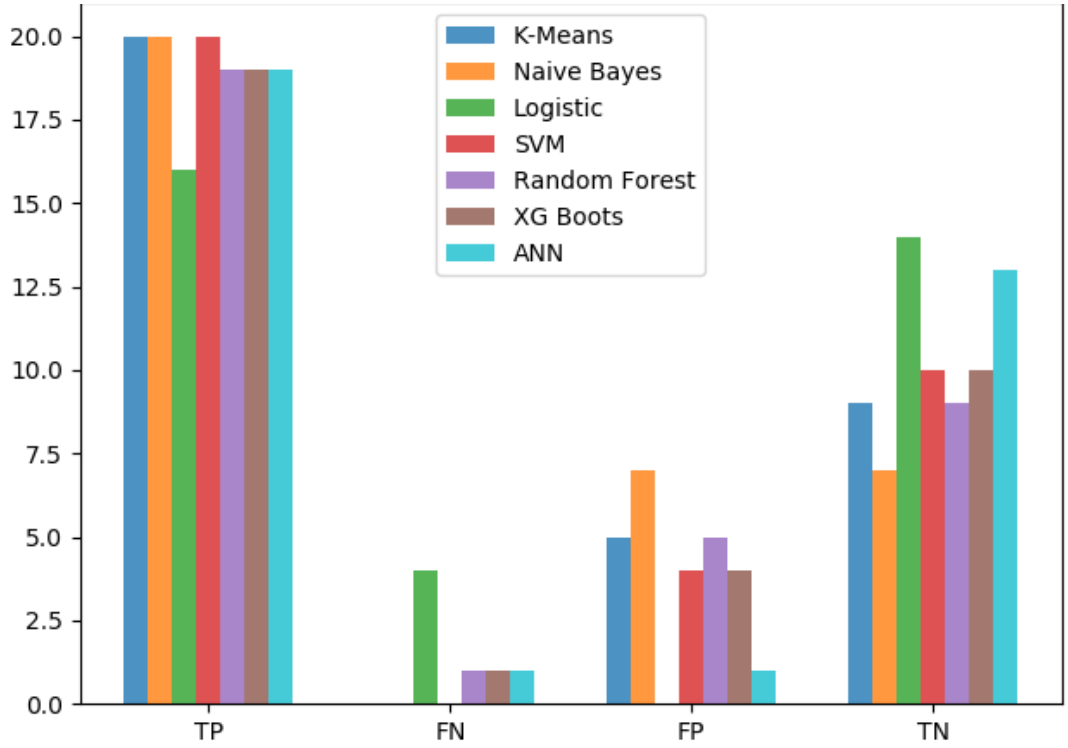




3 Kết Quả

3.2 Confusion matrix

- Naive Bayes: 7 trường hợp rơi vào false positive - Độ chính xác kém nhất
- ANN: 1 trường hợp rơi vào false negative và 1 trường hợp rơi vào false positive - Độ chính xác cao nhất.



	TP	FN	FP	TN
K-Means	20	0	5	9
Naive Bayes	20	0	7	7
Logistic	16	4	0	14
SVM	20	0	4	10
Random Forest	19	1	5	9
XG Boots	19	1	4	10
ANN	19	1	1	13

4 Kết Luận

- Độ chính xác của mô hình ANN cao nhất khi so sánh với tất cả các mô hình máy học còn lại trên tất cả các phương pháp đánh giá độ chính xác.
- Đối với dữ liệu hiện có, số mẫu khá nhỏ so với số biến, do đó độ chính xác không đạt quá cao: 94,1%.
- Sự giảm chiều dữ liệu hiệu quả (từ 7129 biến rút gọn thành 4 thànhphần chính) sẽ giúp cho tốc độ huấn luyện trên mô hình ANN được cải thiện rất nhiều, đặc biệt đối với dữ liệu lớn.

5 Hỏi Đáp

Cảm ơn cô giáo và các bạn đã theo dõi!